DOCUMENT RESUME

ED 218 323 TM 820 363

AUTHOR Modu, Christopher C.
TITLE The Robustness of Latent Trait Model for Achievement Test Score Equating.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY College Entrance Examination Board, New York, N.Y.
PUB DATE Mar 82
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *Equated Scores; Higher Education; *Latent Trait Theory; Testing Problems
IDENTIFIERS Advanced Placement Examinations (CEEB); Test Equivalence; Three Parameter Model; Unidimensional Scaling

ABSTRACT

The equating of scores on alternate forms of different achievement tests through the use of the three-parameter latent trait model, item-response theory (IRT) equating, was compared with the results of score equatings based on conventional linear and curvilinear equating models. Ten equatings were completed for pairs of alternate forms of the Advanced Placement Program, which measures different content areas and traits in each subject area. It was found that despite the apparent violation of the unidimensionality assumption, the equating results obtained through the IRT equating model were found to be in agreement with those of the conventional equating models. By demonstrating that the IRT equating results parallel those of the simpler, less costly, conventional methods, it has been shown that it is still possible to equate scores on non-parallel tests under conditions which make conventional equating inapplicable. (Author/PN)

THE ROBUSTNESS OF LATENT TRAIT MODEL FOR ACHIEVEMENT TEST SCORE EQUATING

by

Christopher C. Modu
Educational Testing Service
Princeton, N. J.

Paper presented at
The American Educational Research Association
Annual Meeting, New York City
March 1982

Purpose of the Study[1]

The equating of scores on non-parallel forms of a test through the application of the three-parameter latent trait model (Lord, 1980), hereafter referred to as item-response theory (IRT) equating, has been shown (Marco, Petersen and Stewart, 1980; Petersen, Cook and Stocking, 1981) to yield at least as accurate, and in some instances, more accurate results than those of conventional linear and curvilinear equating models (Angoff, 1971) for the College Board Scholastic Aptitude Test (SAT). In the second study, Petersen et al. investigated the drift in SAT score scale by comparing the results obtained from the conventional and IRT equating methods. Their study design involved the equating of a test to itself in a circular chain through a series of links (e.g., a ⟶ b ⟶ c ⟶ d ⟶ e ⟶ f ⟶ a) in which each new test is equated to a previous one through an anchor test common to the adjacent pair of tests being equated. The extent of the scale drift was then determined as the difference between the scaled-score conversions for each raw score on test _a_ at the start and at the end of the circular chain. They concluded from their results that the smallest scale drift occurred under the IRT equating method.

If IRT equating works for the SAT, can it also work for achievement tests? Achievement tests, in general, may not satisfy the assumption of unidimensionality which underlies the use of latent trait models. Therefore, the primary purpose of this investigation is to explore the extent to which IRT equating results parallel those of conventional equating methods under conditions which probably violate the unidimensionality assumption.

---

[1]This investigation was supported by the College Entrance Examination Board through its testing and research programs. The author wishes to thank Martha Stocking for assisting with the LOGIST runs and to Samuel Livingston for his helpful comments on the draft of this report.

Another reason for exploring the feasibility of IRT equating for different types of achievement tests is that, under the current test-disclosure environment, it may not even be possible to locate a single previous edition with a sufficient number of items in common with a new edition to allow for the use of conventional equating models. But IRT equating requires only that a sufficient number of items on a new test edition will have been calibrated and placed on a common ability scale. Therefore, IRT equating could still be accomplished even if the calibrated items on the new test edition had been drawn from several previous editions.

Design of the Study

The multiple-choice sections of 11 achievement examinations administered in the College Board Advanced Placement program were used for the study. Except for two 45-minute examinations in Physics C (Mechanics, and Electricity and Magnetism), the remaining nine were made up of 75- to 90-minute examinations.

The equated scores on two editions, A and B, of each achievement examination were determined by three equating methods: the conventional linear and equipercentile equating methods described by Angoff (1971, pp. 568-83) and the three-parameter IRT equating method. For a given test A score, the equated test B scores obtained under the two conventional equating methods were then compared with the corresponding test B score obtained under the IRT equating method.

All three equating procedures used internal anchor tests ranging from 14 to 30 questions. For the IRT equating, the internal anchor test was used to transform the item parameters for each total test to a common ability scale.

The program LOGIST (Wood, Wingersky & Lord, 1976; Wood & Lord, 1976) was used to obtain the item parameter estimates from which the true-score equating of raw scores on tests A and B was accomplished.

Although it would have strengthened the study to confirm by factor analytic methods that the exams used in the study are not unidimensional, the diversity of the content areas encompassed by some of those exams leaves little doubt about their being far from unidimensional. The 120-item biology exam, for example, was made up of questions in three specific content areas: organismal, molecular and populational biology, each area testing knowledge of facts, principles and processes of biology, understanding the means by which biological information is collected, how it is interpreted, and how one formulates hypotheses from available data and makes further predictions. The chemistry exam contained questions on structure of matter, states of matter, chemical reactions, and descriptive chemistry. The questions dealt with understanding and application of principles or calculations or observations and conclusions in experimental situations, etc. The physics exam tested knowledge of physics and the ability to interpret and apply the knowledge both qualitatively and quantitatively, determine directions of vectors or paths of particles or light rays, draw or interpret diagrams, account for observed phenomena, interpret or express physical relationships in graphical forms, manipulate equations and solve problems. The foreign language exams, comprising listening, reading, writing and speaking components, tested the ability to comprehend formal and informal spoken language, the acquisition of vocabulary and a grasp of structure as well as the ability to express ideas orally with accuracy and fluency.

The conventional and the IRT equatings used independent representative samples from the total candidate groups for tests A and B. Part of the reason for not using the same sample is that equating was done long after the operational program administration. Also, since the cost of LOGIST is directly related to sample size, it was necessary to reduce the size of some of the IRT equating samples.

Table 1 shows the examinations used for the equatings, the total number of items in the two editions, A and B, of each examination, the number of common items, and the number of students in equating samples and the total candidate group for each test edition.

Equating Results

Tables 2.a.—2.e. show the equivalent scores on Form B for each of the three equating methods for selected score points on Form A. The linear conversion parameters for transforming the Form A scores to their equivalent Form B scores are indicated at the bottom of the tabulations for each examination. These parameters were derived from the Tucker observed-score linear equating model in preference to the calculations which had also been obtained by applying the Levine equating model (Angoff, 1971). The decision rule as to which of the two linear model's equating results should be used for score reporting depends on the differences in ability level between the groups that took the test editions being equated as well as on the degree of parallelism between the tests. For non-parallel tests administered to groups that are not widely discrepant in ability (as is usually the case for the caliber of total-group candidates for the Advanced Placement program) the Tucker linear model was indicated for score reporting.

Tables 2.a.—2.e. show that the results of the different equating methods are in very close agreement, not differing by more than one point, except at the two extremities of each scale where score equivalences are not usually as accurate because of the scarcity of data at those score levels. These observations are further confirmed by the graphs of the equated scores in Figures A-K. The close agreement between the results of the three equating methods, particularly those of the IRT and equipercentile methods, confirms that the IRT equating method can be used to generate scores that are equivalent to those of conventional equating methods.

Conclusion

Although the unidimensionality of the tests used in this equating experiment was not directly tested, the wide diversity of their content specifications, the behavioral aspects of the skills and abilities tested as well as the multidimensionality of corresponding tests for similar ability groups clearly suggest that one could not safely assume that the tests used for this study are unidimensional. Despite the apparent violation of that assumption, the equating results obtained through the IRT equating model were found to be in agreement with those of the conventional equating models. The application of factor analytic procedures to demonstrate the multidimensionality of the tests used in this investigation would have strengthened the study. It is, however, recommended that a replication of the present study include a design for establishing the extent of scale drift under each of the three equating models by equating a test to itself through a series of intermediate tests in cyclical chain link.

By demonstrating that the IRT equating results parallel those of the simpler, less costly, conventional methods, it has been shown that it is still possible to equate scores on non-parallel tests under conditions which make conventional equating inapplicable. Such a situation will arise when anchor items embedded in a new test cannot be drawn from a single previous edition but from several previous editions containing calibrated items.

REFERENCES

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Lord, F. M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, N. J.: Earlbaum, 1980.

Marco, G. L., Petersen, N. S., and Stewart, E. E. "A Test of the Adequacy of Curvilinear Equating Models." In David J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Computerized Adaptive Testing Laboratory, Department of Psychology, University of Minnesota, September 1980.

Petersen, N. S., Cook, L. L., and Stocking, M. "IRT Versus conventional equating methods: A comparative study of scale stability." Educational Testing Service, Princeton, N. J. (Paper presented at the annual meeting of AERA, Los Angeles, 1981).

Wood, R L, & Lord, F. M. A user's guide to LOGIST. Research Memorandum 76-4. Princeton, N. J.: Educational Testing Service, 1976.

Wood, R L, Wingersky, M. S., & Lord, F. M. LOGIST - A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, N. J.: Educational Testing Service, 1976.

Table 1

Tests and Equating Samples

EXAMINATION	No. of Common Equating Items	CONVENTIONAL EQUATING: Number of Students — Old Form (A) Sample	New Form (B) Sample	IRT EQUATING: Number of Students — Old Form (A) Sample	Old Form (A) (Total Grp)	New Form (B) Sample	New Form (B) (Total Grp)
1. AMERICAN HISTORY A(100)→B(79)@	21	4,901	4,847	1,782	(21,080)	3,114	(28,079)
2. BIOLOGY A(120)→B(120)	30	4,843	5,422	1,614	(10,377)	3,165	(12,782)
3. CHEMISTRY A(80)→B(80)	20	6,084	3,219	3,048	(6,188)	2,694	(8,084)
4. EUROPEAN HISTORY A(110)→B(90)	21	5,799	3,245	2,899	(5,871)	3,982	(7,965)
5. FRENCH LANGUAGE A(100)→B(100)	23	1,550*	1,692*	1,533*	(1,574)	2,775*	(2,775)
6. MATH: CALCULUS AB A(45)→B(45)	15	3,277	2,949	1,869	(13,885)	3,092	(15,581)
7. MATH: CALCULUS BC A(45)→B(45)	15	6,524	2,971	3,259	(6,616)	3,850	(7,712)
8. PHYSICS B A(68)→B(70)	23	1,605	1,647	1,604	(1,610)	2,385	(2,385)
9. PHYSICS C (MECH.) A(35)→B(35)	14	1,462	1,402	1,460	(1,489)	2,096	(2,099)
10. PHYSICS C (E&M) A(35)→B(35)	14	1,220	1,057	1,222	(1,240)	1,669	(1,674)
11. SPANISH LANGUAGE A(90)→B(90)	27	1,056*	1,249*	1,040*	(1,066)	2,805*	(2,805)

@Number of questions in each test form is indicated in parentheses, e.g., 100 in Form A and 79 in Form B for American History.

*Available "standard" group, i.e., those candidates who are non native-speakers and who have spent less than 1 month in a French- or Spanish-speaking country.

Table 2.a.

Comparison of Raw to Raw Score Conversions
Obtained from Conventional and IRT Equating Methods

AMERICAN HISTORY

FORM A (100 max.)	EQUIVALENT FORM B SCORE (MAX. POSS. 79) IRT	EQUIPER-CENTILE	LINEAR* (TUCKER)
100	79	79	73
90	69	69	66
80	60	60	58
74	55	54	54
70	51	51	51
60	44	43	44
59	43	43	43
50	36	36	36
45	33	32	33
40	29	29	29
30	23	22	22
22	17	17	16
16	13	13	11
15	12	12	11
10	9	9	7
6	6	6	4
0	2	3	0

*FORM B = .7327(A) - 0.3762

EUROPEAN HISTORY

FORM A (110 max.)	EQUIVALENT FORM B SCORE (MAX. POSS. 90) IRT	EQUIPER-CENTILE	LINEAR* (TUCKER)
110	90		90
100	82	83	83
90	73	74	74
80	65	66	66
72	59	60	60
60	50	51	50
50	42	42	41
40	35	34	33
[illegible]	31	30	30
30	26	25	25
24	21	20	20
20	17	16	17
10	8	7	8
5	4	1	4
0	0	0(-1)	0

*FORM B = 0.8250(A) + 0.1954

Table 2.b.

Comparison of Raw to Raw Score Conversions
Obtained from Conventional and IRT Equating Methods

BIOLOGY

FORM A (120 max.)	EQUIVALENT FORM B SCORE (MAX. POSS. 120) IRT	EQUIPER-CENTILE	LINEAR* (TUCKER)
120	120	120	118
110	110	110	108
100	99	98	98
90	87	88	88
85	82	83	83
70	67	68	68
60	57	58	58
47	44	44	44
40	38	38	37
31	29	28	28
20	19	17	17
10	10	8	7
0	1	0	0(-3)

*FORM B = 1.0143(A) - 3.2529

CHEMISTRY

FORM A (80 max.)	EQUIVALENT FORM B SCORE (MAX. POSS. 80) IRT	EQUIPER-CENTILE	LINEAR* (TUCKER)
80	80	75	73
75	73	71	68
70	67	65	63
60	55	54	54
59	54	53	53
50	45	45	45
48	43	43	43
40	35	35	35
32	28	28	28
25	21	21	21
22	19	19	18
20	17	17	16
10	8	8	7
5	4	4	2
0	0	0	0(-2)

*FORM B = .9375(A) - 2.3118

Table 2.c.

Comparison of Raw to Raw Score Conversions
Obtained from Conventional and IRT Equating Methods

CALCULUS AB

EQUIVALENT FORM B SCORE
(MAX. POSS. 45)

FORM A (45 max.)	IRT	EQUIPER-CENTILE	LINEAR* (TUCKER)
45	45	44	42
40	39	38	37
36	34	37	33
29	26	26	26
25	22	22	23
20	17	17	18
15	12	12	13
12	10	10	10
10	8	9	8
5	4	5	3
0	0	0	0(-2)

*FORM B = 0.9819(A) - 1.9861

CALCULUS BC

FORM A (45 max.)	IRT	EQUIPER-CENTILE	LINEAR* (TUCKER)
45	45	45	44
40	40	40	39
36	35	35	35
30	29	29	28
25	23	23	23
20	18	18	18
18	16	16	16
13	11	10	10
10	8	7	7
5	3	3	2
0	0(-1)	0	0(-3)

*FORM B = 1.0556(A) - 3.2515

Table 2.d.

Comparison of Raw to Raw Score Conversions
Obtained from Conventional and IRT Equating Methods

PHYSICS-B

	EQUIVALENT FORM B SCORE (MAX. POSS. 70)		
FORM A (68 Max.)	IRT	EQUIPER-CENTILE	LINEAR* (TUCKER)
68	70	66	70
60	62	66	62
50	52	59	52
43	45	46	44
40	42	42	41
33	35	35	34
20	21	21	21
16	17	17	17
10	11	11	11
5	6	6	6
0	1	0	1

*FORM B = 1.0207(A) + .5478

PHYSICS (MECHANICS)

	EQUIVALENT FORM B SCORE (MAX. POSS. 35)		
FORM A (35 max.)	IRT	EQUIPER-CENTILE	LINEAR* (TUCKER)
35	35	35	34
30	29	29	29
25	24	24	24
21	20	20	20
14	13	14	13
10	9	9	9
6	6	6	6
5	5	5	5
0	0	0	0

*FORM B = 0.9857(A) - 0.3743

PHYSICS (ELEC. & MAGNETISM)*

35	35	34	35
30	30	30	30
25	25	25	25
16	16	16	16
11	11	11	11
7	7	6	7
4	4	4	4
1	1	1	0 (.47)

*FORM B = 1.0163(A) - 0.5462

Table 2.e.

Comparison of Raw to Raw Score Conversions
Obtained from Conventional and IRT Equating Methods

FRENCH LANGUAGE

FORM A (100 Max.)	EQUIVALENT FORM B SCORE (MAX. POSS, 100) IRT	EQUIPER-CENTILE	LINEAR* (TUCKER)
100	100		100(108)
90	92	93	93
80	84	84	82
73	77	74	75
65	68	66	66
57	58	58	58
50	50	50	50
39	38	37	38
35	33	34	34
30	28	29	28
26	24	24	24
20	19	19	17
10	10	8	7
5	5	4	1
0	0	0	0(-4)

*FORM B = 1.0844(A) - 4.2614

SPANISH LANGUAGE

FORM A (90 max.)	EQUIVALENT FORM B SCORE (MAX. POSS. 90) IRT	EQUIPER-CENTILE	LINEAR* (TUCKER)
90	89		88
85	84		83
80	80	79	78
70	69	70	67
60	58	58	57
54	51	51	50
50	46	46	46
45	40	41	41
40	35	35	36
32	27	27	27
25	20	19	20
20	16	14	14
10	8	6	4
0	1	0	0(-7)

*FORM B = 1.0530(A) - 6.5958

COMPARISON OF SCORE CONVERSIONS DERIVED FROM THREE EQUATING METHODS FOR AMERICAN HISTORY

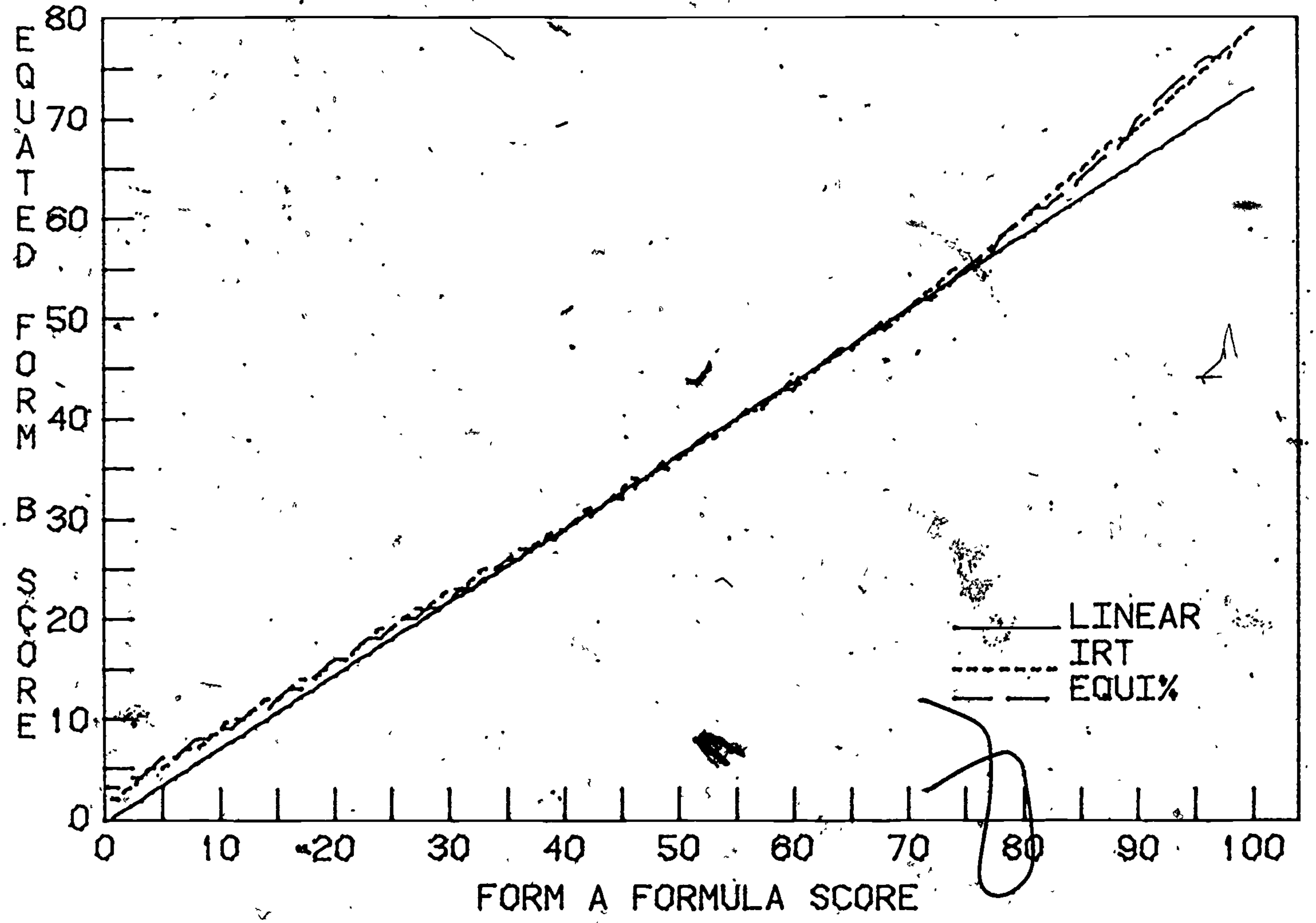


Figure A

COMPARISON OF SCORE CONVERSIONS DERIVED FROM THREE EQUATING METHODS FOR EUROPEAN HISTORY

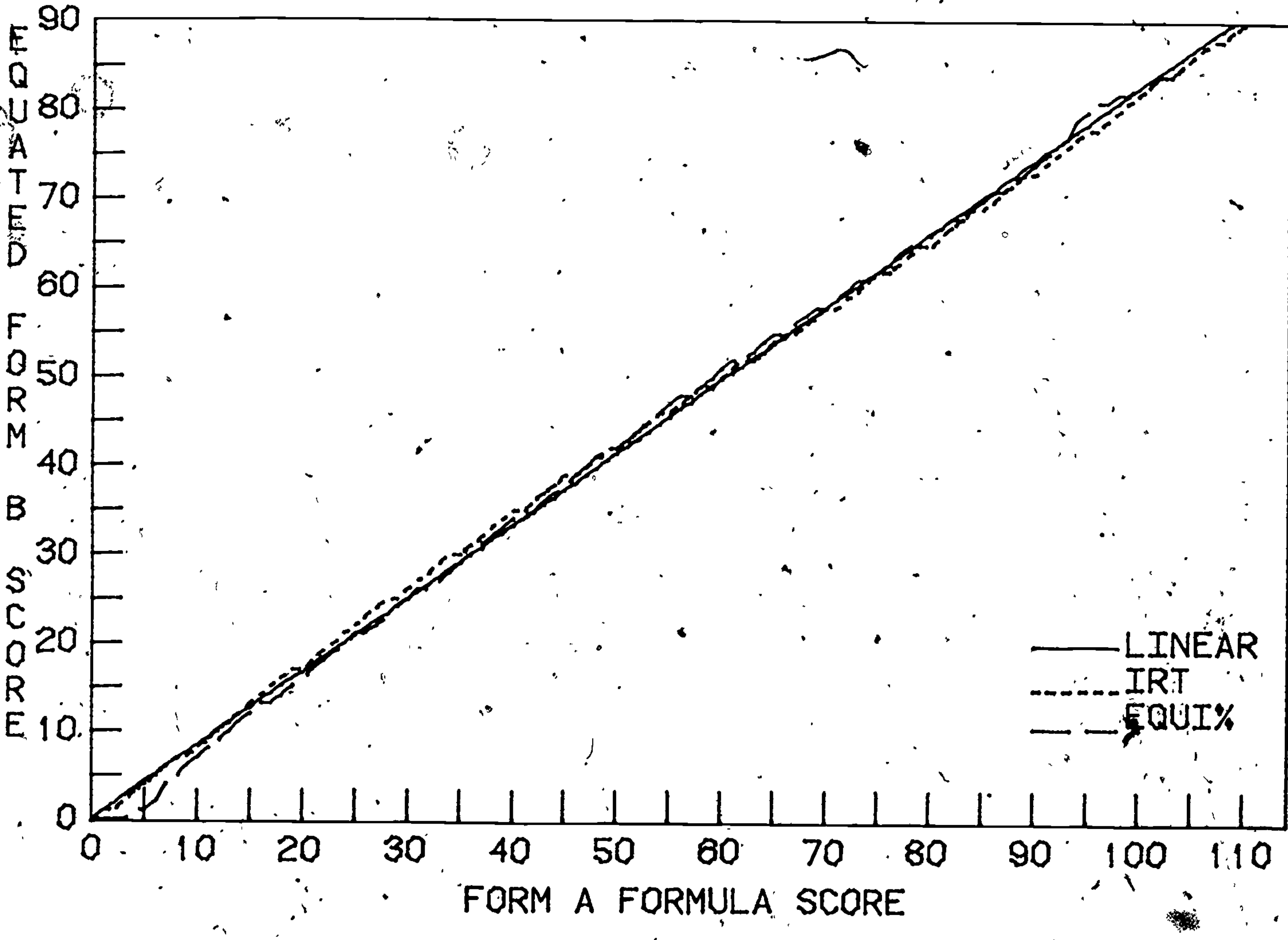
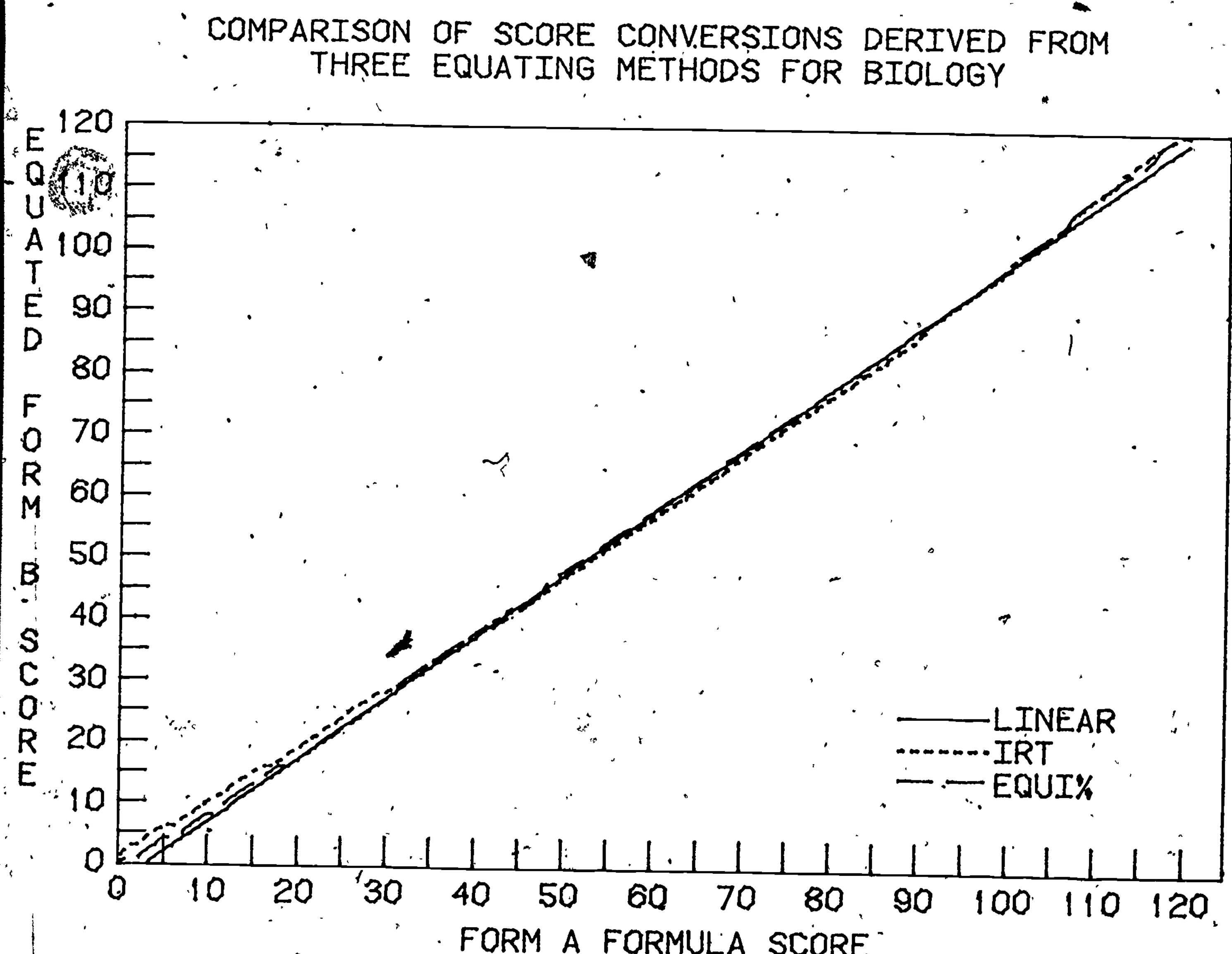
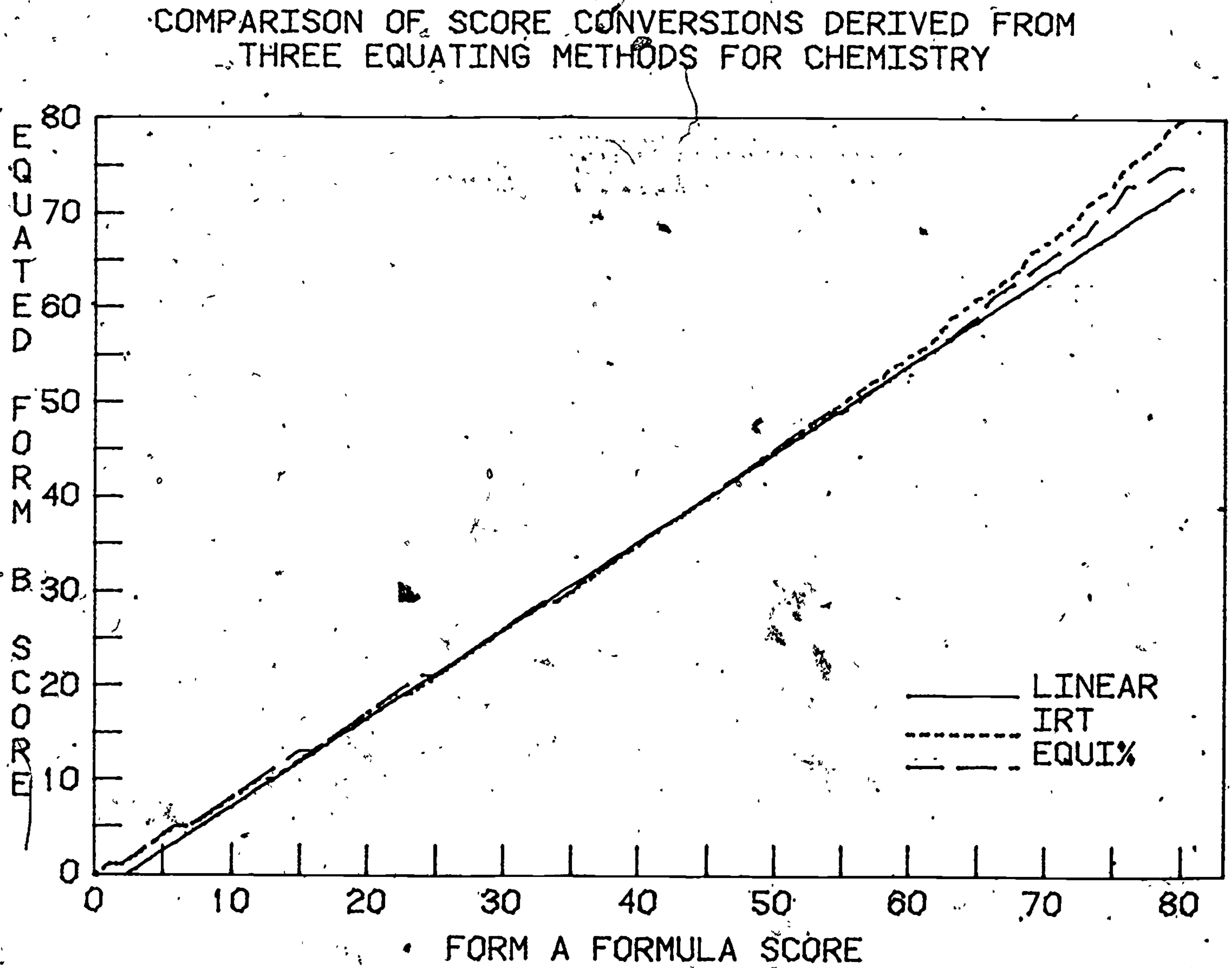


Figure B



COMPARISON OF SCORE CONVERSIONS DERIVED FROM THREE EQUATING METHODS FOR BIOLOGY
EQUATED FORM B SCORE
0
10
20
30
40
50
60
70
80
90
100
110
120
FORM A FORMULA SCORE
LINEAR
IRT
EQUI%

Figure C



COMPARISON OF SCORE CONVERSIONS DERIVED FROM THREE EQUATING METHODS FOR CHEMISTRY
EQUATED FORM B SCORE
FORM A FORMULA SCORE
0
10
20
30
40
50
60
70
80
LINEAR
IRT
EQUI%

Figure D

COMPARISON OF SCORE CONVERSIONS DERIVED FROM THREE EQUATING METHODS FOR CALCULUS AB

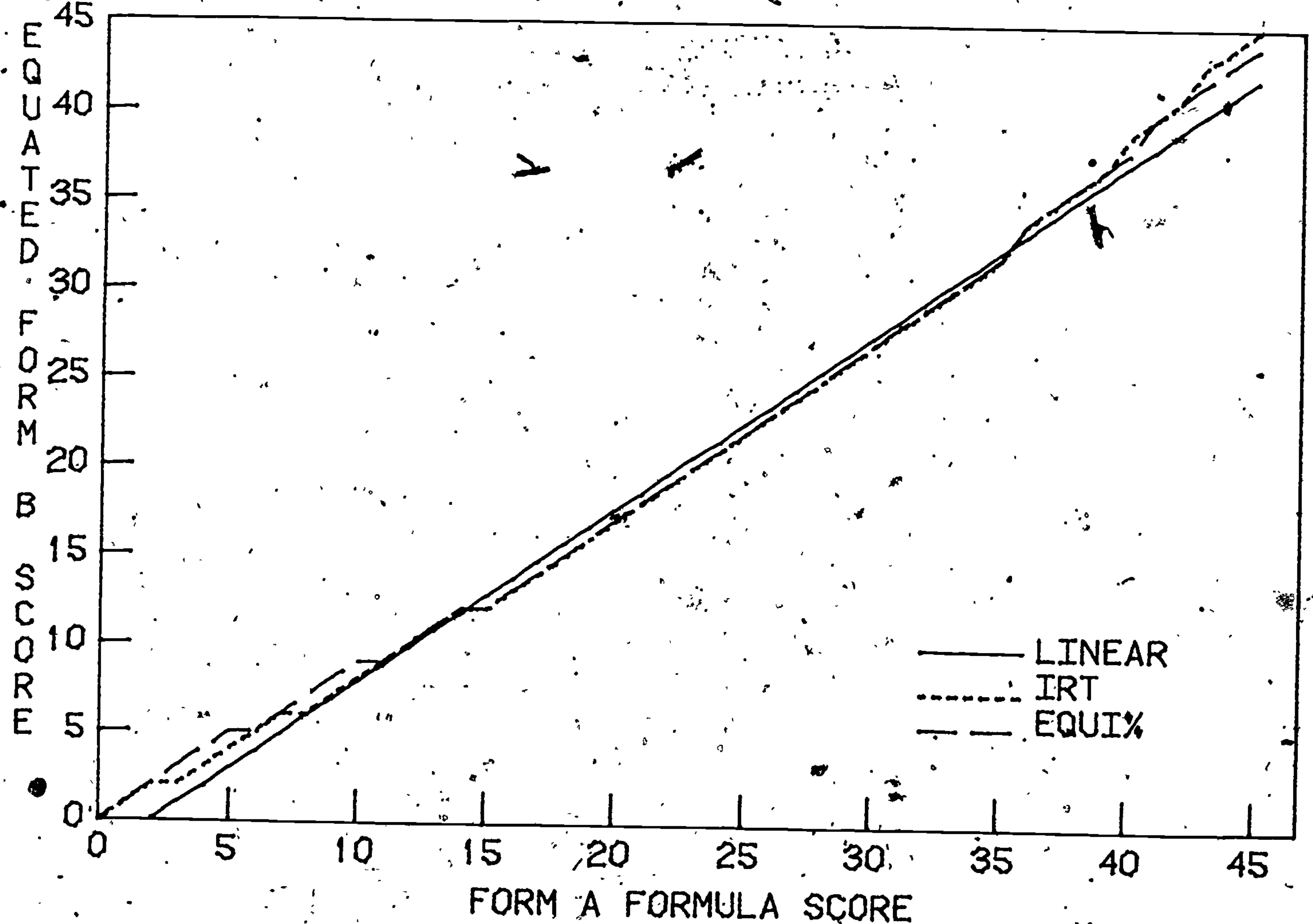


Figure E

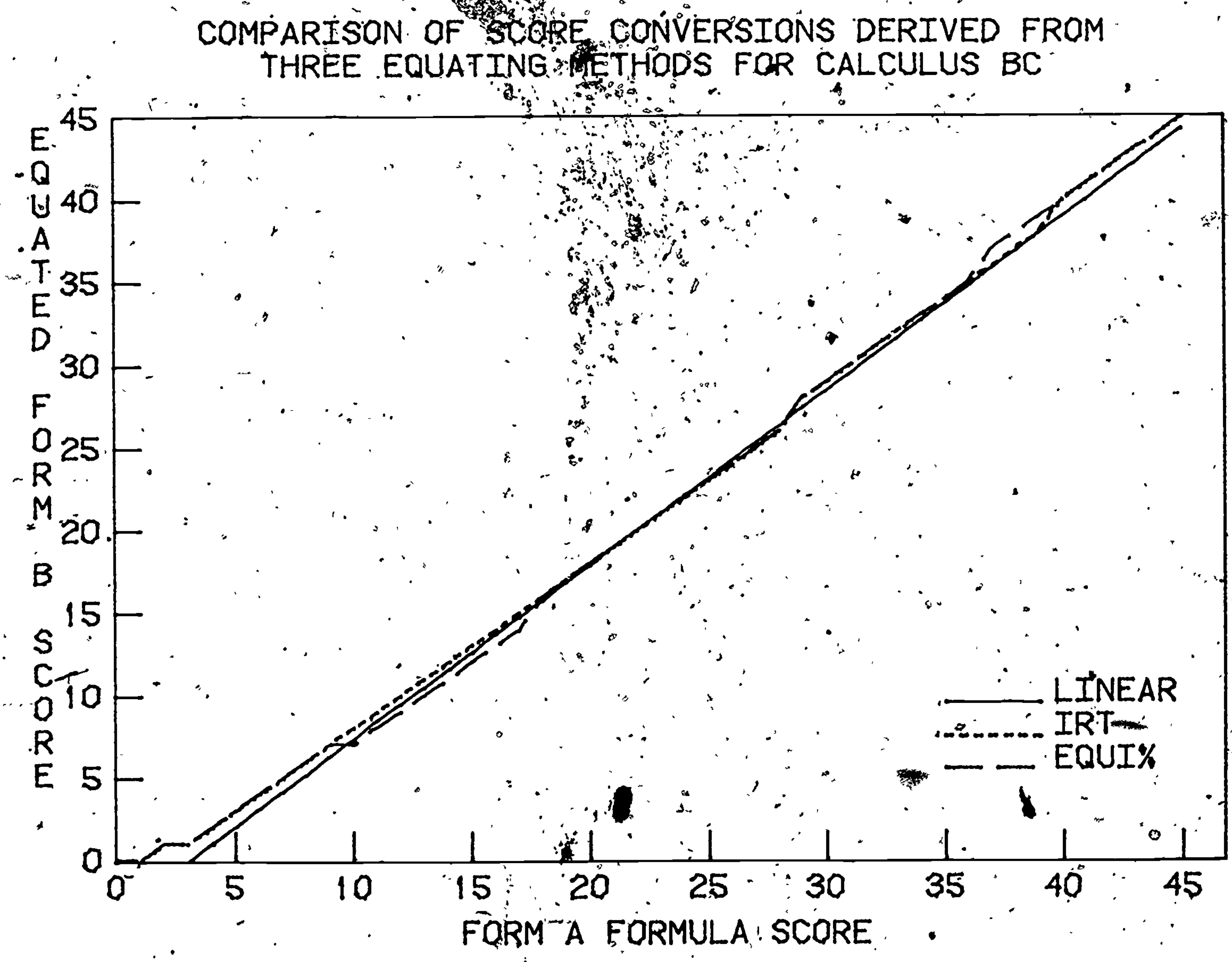

COMPARISON OF SCORE CONVERSIONS DERIVED FROM
THREE EQUATING METHODS FOR CALCULUS BC
EQUATED FORM B SCORE
0
5
10
15
20
25
30
35
40
45
FORM A FORMULA SCORE
LINEAR
IRT
EQUI%

Figure F

COMPARISON OF SCORE CONVERSIONS DERIVED FROM THREE EQUATING METHODS FOR PHYSICS B

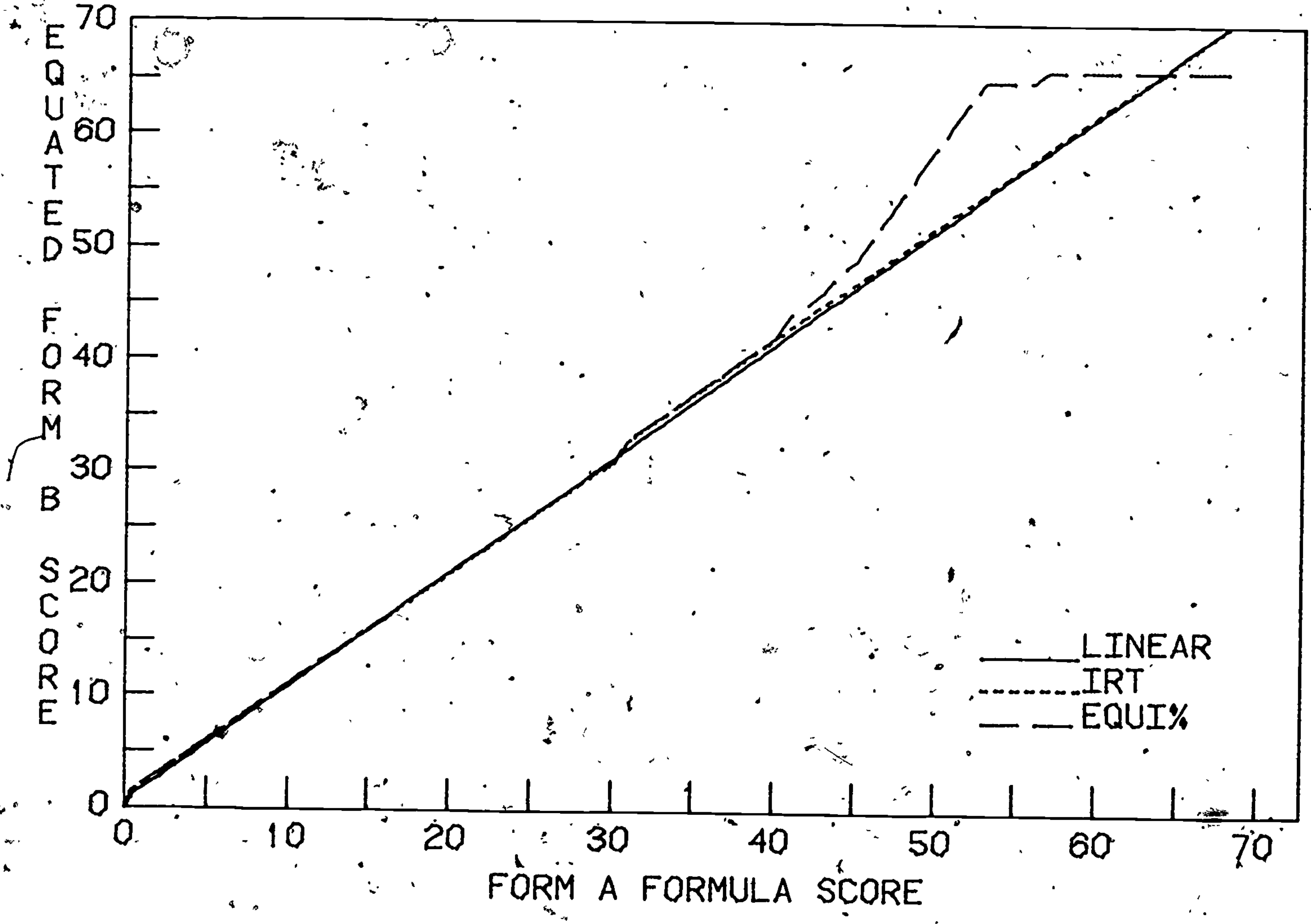


Figure G

COMPARISON OF SCORE CONVERSIONS DERIVED FROM THREE EQUATING METHODS FOR PHYSICS - MECHANICS

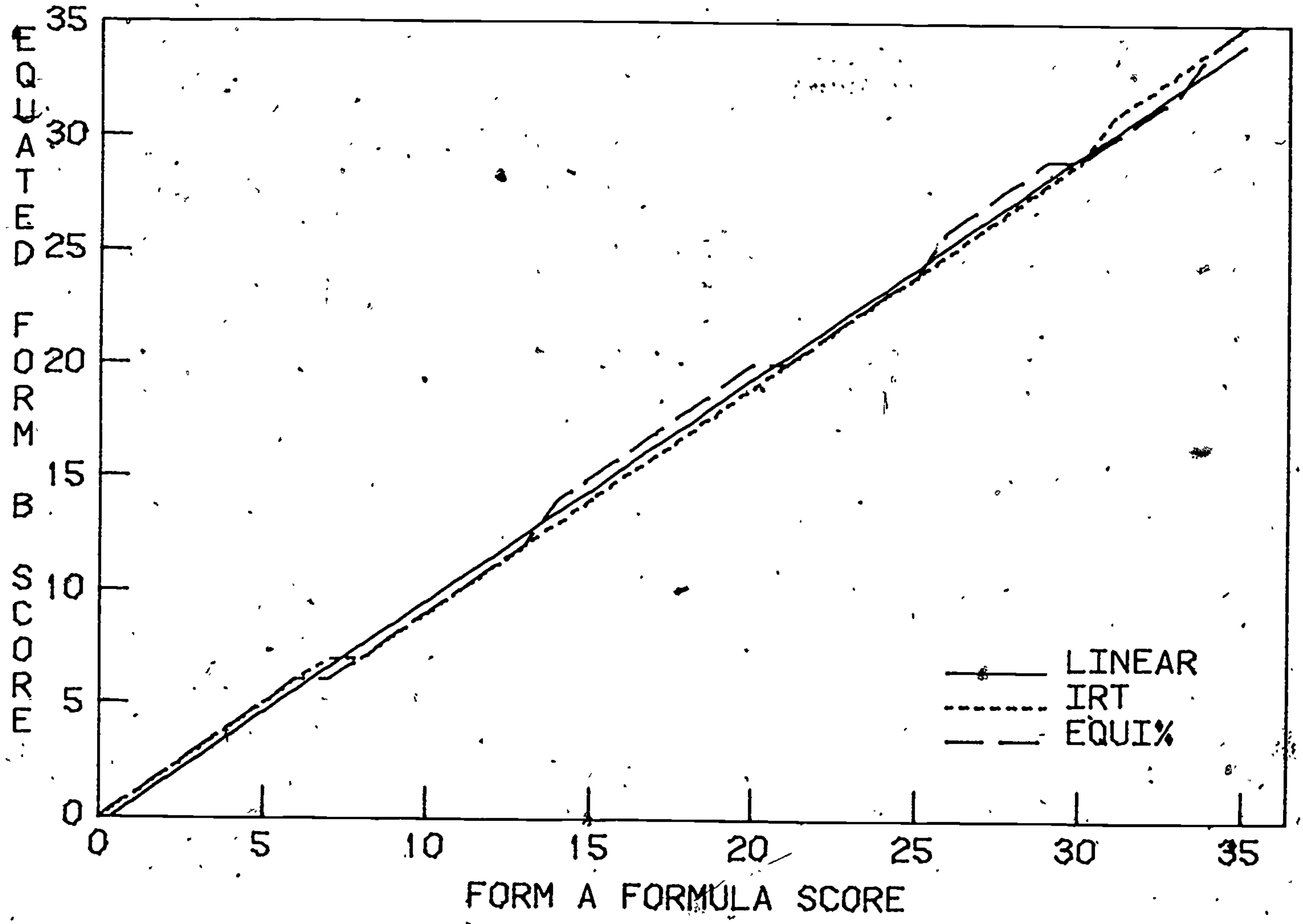


Figure H

COMPARISON OF SCORE CONVERSIONS DERIVED FROM THREE EQUATING METHODS FOR PHYSICS - ELEC. & MAGNETISM

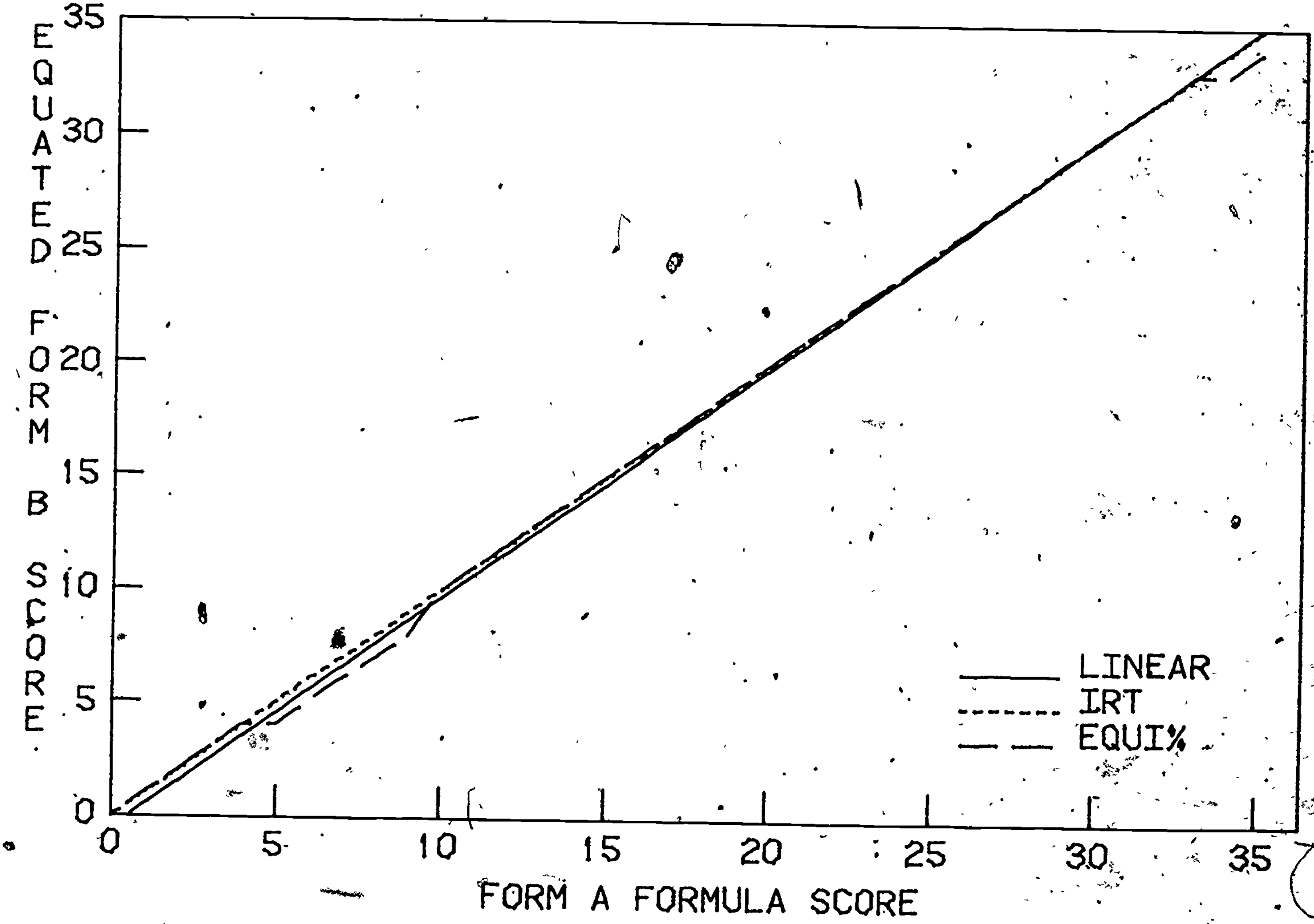


Figure I

COMPARISON OF SCORE CONVERSIONS DERIVED FROM THREE EQUATING METHODS FOR FRENCH LANGUAGE

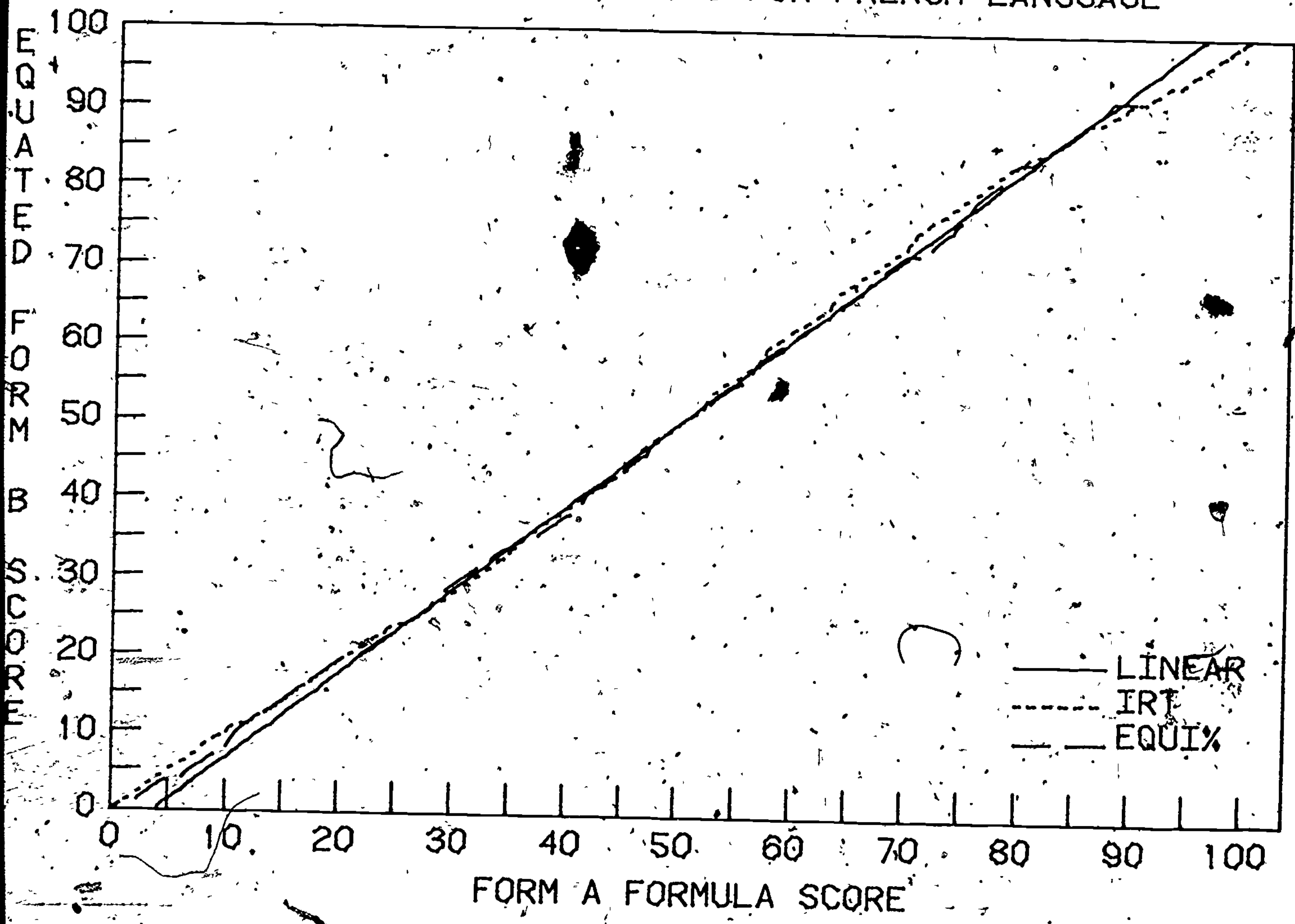


Figure J

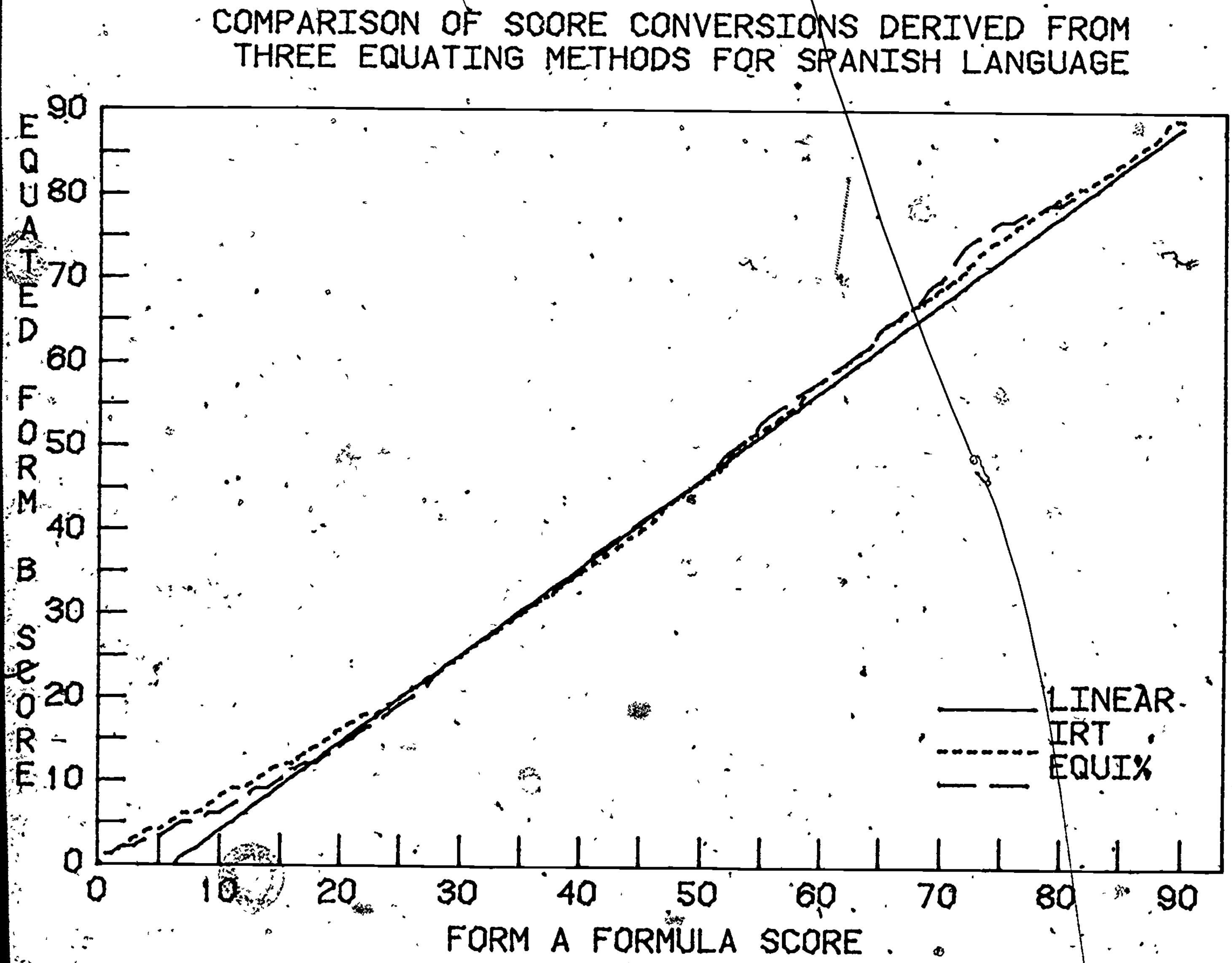

COMPARISON OF SCORE CONVERSIONS DERIVED FROM
THREE EQUATING METHODS FOR SPANISH LANGUAGE
EQUATED FORM B SCORE
0
10
20
30
40
50
60
70
80
90
FORM A FORMULA SCORE
LINEAR
IRT
EQUI%

Figure K